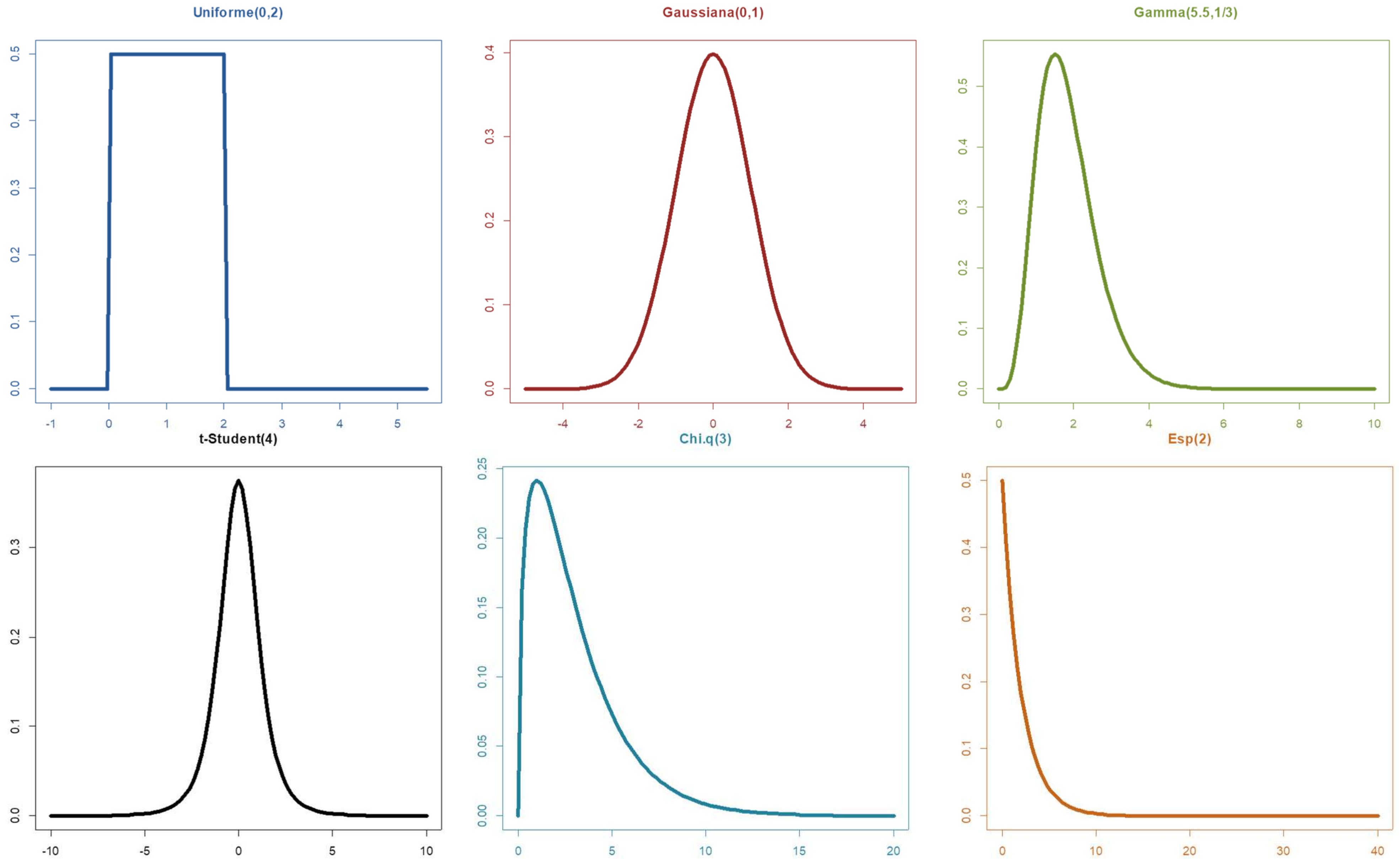
STATISTICA

I modelli probabilistici, III

NON E' CHE ESISTA SOLO LA GAUSSIANA!!!



… MA E' TANTO, TANTO COMODA …

$$X \sim N(\mu, \sigma^2) \Longrightarrow \qquad X' = aX + b \sim \boldsymbol{N}(a\mu + b, a^2\sigma^2)$$

$$X_1 \sim N(\mu_1, {\sigma_1}^2)\ e\ \ X_2 \sim N(\mu_2, {\sigma_2}^2)\ \boldsymbol{indipendenti} \ \Longrightarrow$$

$$X_1 + X_2 \sim \boldsymbol{N}(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$$

… MA E' TANTO, TANTO COMODA …

$$X \sim N(\mu, \sigma^2) \Longrightarrow \qquad X' = aX + b \sim \boldsymbol{N}(a\mu + b, a^2\sigma^2)$$

$$X_1 \sim N(\mu_1, {\sigma_1}^2) \; e \;\; X_2 \sim N(\mu_2, {\sigma_2}^2) \; \boldsymbol{indipendenti} \;\Longrightarrow$$

$$X_1 + X_2 \sim \boldsymbol{N}(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$$

$$X_1 \sim N(\mu, \sigma^2) \; e \;\; X_2 \sim N(\mu, \sigma^2) \; \boldsymbol{indipendenti \; (e \; identiche)} \Longrightarrow$$

$$X_1 + X_2 \sim \boldsymbol{N}(2\mu, 2\sigma^2)$$

$$\frac{X_1 + X_2}{2} \sim \boldsymbol{N}\left(\mu, \frac{\sigma^2}{2}\right)$$

… MA E' TANTO, TANTO COMODA …

$$X \sim N(\mu, \sigma^2) \Longrightarrow \qquad X' = aX + b \sim \boldsymbol{N}(a\mu + b, a^2\sigma^2)$$

$$X_1 \sim N(\mu_1, {\sigma_1}^2)\ e\ \ X_2 \sim N(\mu_2, {\sigma_2}^2)\ \boldsymbol{indipendenti} \Longrightarrow$$

$$X_1 + X_2 \sim \boldsymbol{N}(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$$

$$X_1, \dots, X_n\ \ \boldsymbol{indipendenti\ e\ identiche}, X_i \sim N(\mu, \sigma^2) \Longrightarrow$$

$$X_1 + X_2 + \cdots + X_n \sim \boldsymbol{N}(n\mu, n\sigma^2)$$

$$\boxed{\overline{\boldsymbol{X}}_{\boldsymbol{n}} = \frac{\boldsymbol{X_1} + \boldsymbol{X_2} + \cdots + \boldsymbol{X_n}}{\boldsymbol{n}} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\right)}$$

UNA PARENTESI

FONDAMENTALE

$X_1, \dots, X_n$ ***indipendenti e identiche***, $X_i \sim N(\mu, \sigma^2) \Longrightarrow$

$$X_1 + X_2 + \cdots + X_n \sim \boldsymbol{N}(n\mu, n\sigma^2)$$

$$\boldsymbol{\bar{X}_n} = \frac{\boldsymbol{X_1} + \boldsymbol{X_2} + \cdots + \boldsymbol{X_n}}{\boldsymbol{n}} \qquad \boldsymbol{E}(\boldsymbol{\bar{X}_n}) = \mu, \qquad Var(\boldsymbol{\bar{X}_n}) = \frac{\sigma^2}{\boldsymbol{n}}$$

UNA PARENTESI

campione casuale: repliche dell'esperimento, sempre nelle stesse condizioni (estrazioni con reimmissione)

$X_1, \dots, X_n$ ***indipendenti e identiche***, $X_i \sim N(\mu, \sigma^2) \Longrightarrow$

$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$

$$\bar{\boldsymbol{X}}_n = \frac{\boldsymbol{X_1} + \boldsymbol{X_2} + \cdots + \boldsymbol{X_n}}{\boldsymbol{n}} \Rightarrow \boldsymbol{E}(\bar{\boldsymbol{X}}_n) = \mu, \qquad Var(\bar{\boldsymbol{X}}_n) = \frac{\sigma^2}{\boldsymbol{n}}$$

un inciso importante

$X_1\ e\ X_2$ ***indipendenti*** se $\{P(A \cap B) = P(A)P(B)\ \}$

$\forall\ a < b, c < d$ si ha che

$P(a < X_1 \leq b$ e $c < X_2 \leq d) = P(a < X_1 \leq b) \times P(c < X_2 \leq d)$

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 $\mu\mathrm{g/L}$.

Sapendo che la legge impone una soglia massima per l’arsenico di 10 $\mu\mathrm{g/L}$, si può affermare che l'acqua potabile di Milano non è a norma di legge?

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 μg/L.

Sapendo che la legge impone una soglia massima per l’arsenico di 10 μg/L, si può affermare che l'acqua potabile di Milano non è a norma di legge?

$$\bar{x} = \frac{15 + 12.5 + 7.0 + 13.0 + 7.5}{5} = 11\ \mu\text{g/L} > 10\ \mu\text{g/L}$$

è un risultato attendibile per **tutta** l’acqua di Milano?

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 $\mu\mathrm{g/L}$.

Sapendo che la legge impone una soglia massima per l’arsenico di 10 $\mu\mathrm{g/L}$, si può affermare che l'acqua potabile di Milano non è a norma di legge?

Sia X la variabile che descrive la quantità di arsenico in un campione di acqua milanese. Supp. $\boldsymbol{X \sim N(\mu, \sigma^2)}$ con $\boldsymbol{\mu}$ **incognita** e $\sigma = 1.6\ \mu\mathrm{g/L}$ (semplificazione momentanea)

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **qua**[illegible]**senico**, ottenendo i s[illegible]0, 13.0, 7.5 μg/L.

PROVIAMO A RAGIONARE COME CON LA CLINICA DELLE FIGLIE FEMMINE....

Sapendo c[illegible]’arsenico di 10 μg/L, si [illegible] Milano non è a norma di legge?

Sia X la variabile che descrive la quantità di arsenico in un campione di acqua milanese. Supp. $\boldsymbol{X \sim N(\mu, \sigma^2)}$ con $\boldsymbol{\mu}$ **incognita** e $\sigma = 1.6\ \mu\mathrm{g/L}$ (semplificazione momentanea)

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 $\mu\mathrm{g/L}$.

Sapendo che la legge impone una soglia massima per l’arsenico di 10 $\mu\mathrm{g/L}$, si può affermare che l'acqua potabile di Milano non è a norma di legge?

Sia X la variabile che descrive la quantità di arsenico in un campione di acqua milanese. Supp. $\boldsymbol{X \sim N(\mu, \sigma^2)}$ con $\boldsymbol{\mu}$ **incognita** e $\sigma = 1.6\ \mu\mathrm{g/L}$ (semplificazione momentanea)

1. Usiamo **il valore di riferimento di legge per** μ ⇨ $X \sim N(\mathbf{10}, 1.6^2)$

Esempio all'arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 $\mu\mathrm{g/L}$.

Sapendo che la legge impone una soglia massima per l'arsenico di 10 $\mu\mathrm{g/L}$, si può affermare che l'acqua potabile di Milano non è a norma di legge?

Sia X la variabile che descrive la quantità di arsenico in un campione di acqua milanese. Supp. $\boldsymbol{X \sim N(\mu, \sigma^2)}$ con $\boldsymbol{\mu}$ **incognita** e $\sigma = 1.6\ \mu\mathrm{g/L}$ (semplificazione momentanea)

1. Usiamo **il valore di riferimento di legge per $\boldsymbol{\mu}$** ⇨ $X \sim N(\mathbf{10}, 1.6^2)$
2. $X_1, X_2, \cdots, X_5$ variabili ***i.i.d,*** $\boldsymbol{\sim X}$ ⟹ $\boldsymbol{\bar{X}_5 \sim N(10, 1.6^2/5)}$

Esempio all’arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 $\mu\mathrm{g/L}$.

Sapendo che la legge impone una soglia massima per l’arsenico di 10 $\mu\mathrm{g/L}$, si può affermare che l'acqua potabile di Milano non è a norma di legge?

Sia X la variabile che descrive la quantità di arsenico in un campione di acqua milanese. Supp. $\boldsymbol{X \sim N(\mu, \sigma^2)}$ con $\boldsymbol{\mu}$ **incognita** e $\sigma = 1.6\ \mu\mathrm{g/L}$ (semplificazione momentanea)

1. Usiamo **il valore di riferimento di legge per** $\boldsymbol{\mu}$ ⇨ $X \sim N(\mathbf{10}, 1.6^2)$
2. $X_1, X_2, \cdots, X_5$ variabili ***i.i.d.*** $\sim X \Rightarrow \bar{X}_5 \sim \boldsymbol{N(10, 1.6^2/5)}$
3. $P(\bar{X}_5 > 11) = 1 - pnorm(11{,}10, \sqrt{1.6^2/5}) = 0.08112525$

Esempio all'arsenico

Durante le verifiche sulla qualità dell'acqua potabile di Milano si è misurata la **quantità presente nell'acqua di arsenico**, ottenendo i seguenti **valori campionari**: 15, 12.5, 7.0, 13.0, 7.5 μg/L.

Saper ... di 10 ...

SE A MILANO E' RISPETTATO IL LIMITE DI LEGGE, LA PROBABILITA' CHE LA MEDIA DI UN CAMPIONE DI 5 DATI SIA ≥ 11 è DELL'8%

S ... n campione ... μ **incognita** e $\sigma = 1.6\ \mu\text{g/L}$ (semplificazione ...tanea)

1. Usiamo **il valore di riferimento di legge per μ** ➡ $X \sim N(\mathbf{10}, 1.6^2)$
2. $X_1, X_2, \cdots, X_5$ variabili ***i.i.d.*** $\sim X \Rightarrow \bar{X}_5 \sim \boldsymbol{N}\left(\mathbf{10}, \mathbf{1.6^2/5}\right)$
3. $P(\bar{X}_5 > 11) = 1 - pnorm(11, 10, \sqrt{1.6^2/5}) = 0.08112525$

Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

media campionaria: v.c. che *predice* il valore della media aritmetica dei dati nel campione

Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

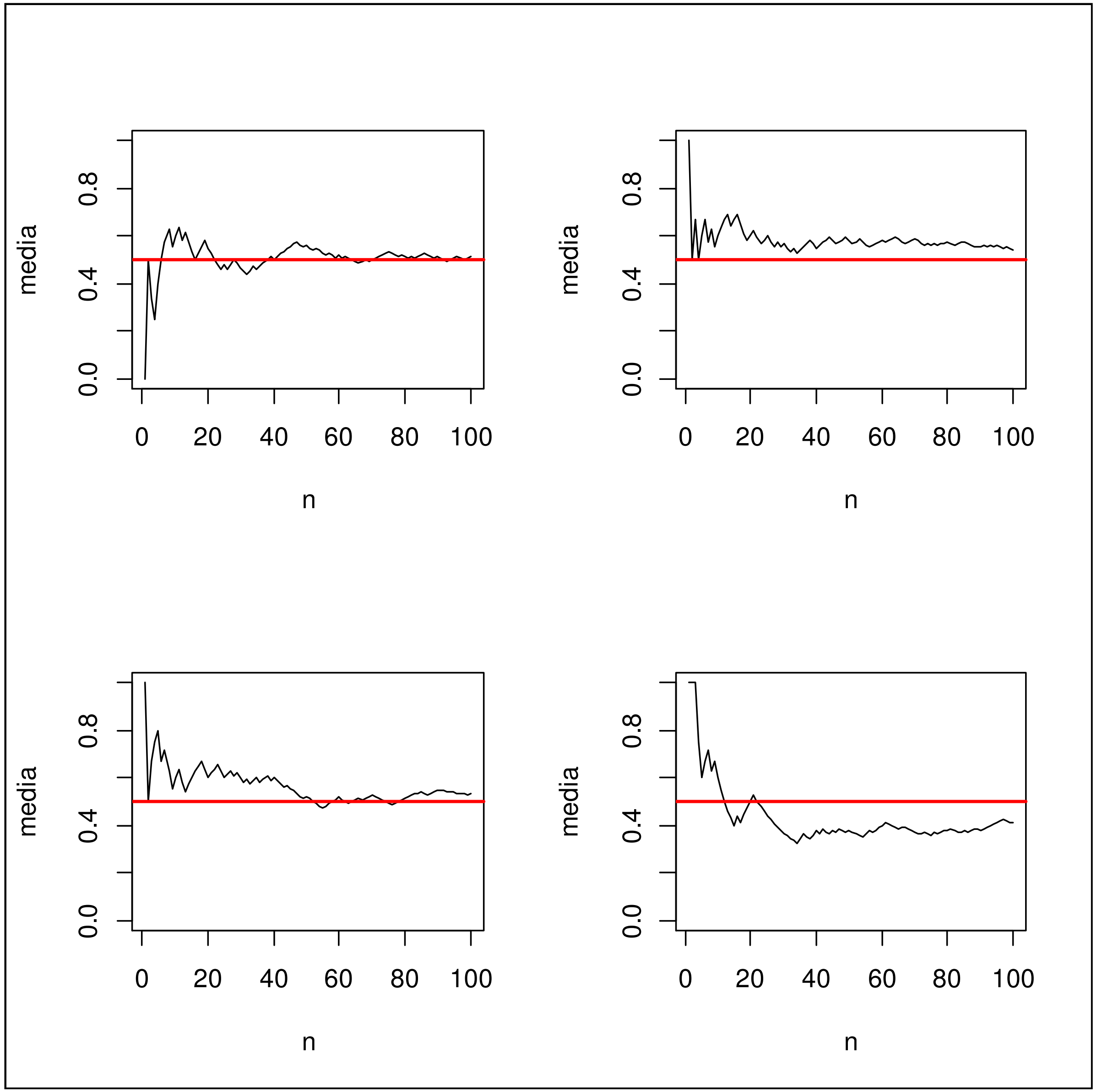
media campionaria: v.c. che *predice* il valore della media aritmetica dei dati nel campione
E' uno *stimatore* della media $E(X_i)$

$X_1, \ldots, X_n, \ldots$ successione di v.a. **i.i.d** con media $E(X_i) = \mu$ finita.

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n} \xrightarrow[n \to \infty]{} \mu \quad \textit{quasi certamente}$$

Legge dei grandi numeri

Facciamo un salto in R



Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

media campionaria

Se $X_i \sim N(\mu, \sigma^2)$ $\Rightarrow$ $\bar{X}_n \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}\right)$ — $\bar{X}_n$ stimatore di $\boldsymbol{\mu}$

Se $X_i \sim b(p)$ $\Rightarrow$ $\bar{X}_n \sim \frac{1}{n} Bin(n, p)$ — $\bar{X}_n$ stimatore di $\boldsymbol{p}$

Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

media campionaria

Se $X_i \sim N(\mu, \sigma^2)$ $\Rightarrow$ $\bar{X}_n \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}\right)$

Se $X_i \sim b(p)$ $\Rightarrow$ $\bar{X}_n \sim \frac{1}{n} Bin(n, p)$

Se $\boldsymbol{np \geq 5}$ & $\boldsymbol{np(1-p) \geq 5}$ $\Rightarrow$ $\bar{X}_n \sim \boldsymbol{N}\left(\boldsymbol{p}, \frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}\right)$

Inferenza per la media

Siano $X_1, X_2, \cdots, X_n$ variabili casuali ***i.i.d***

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

media campionaria

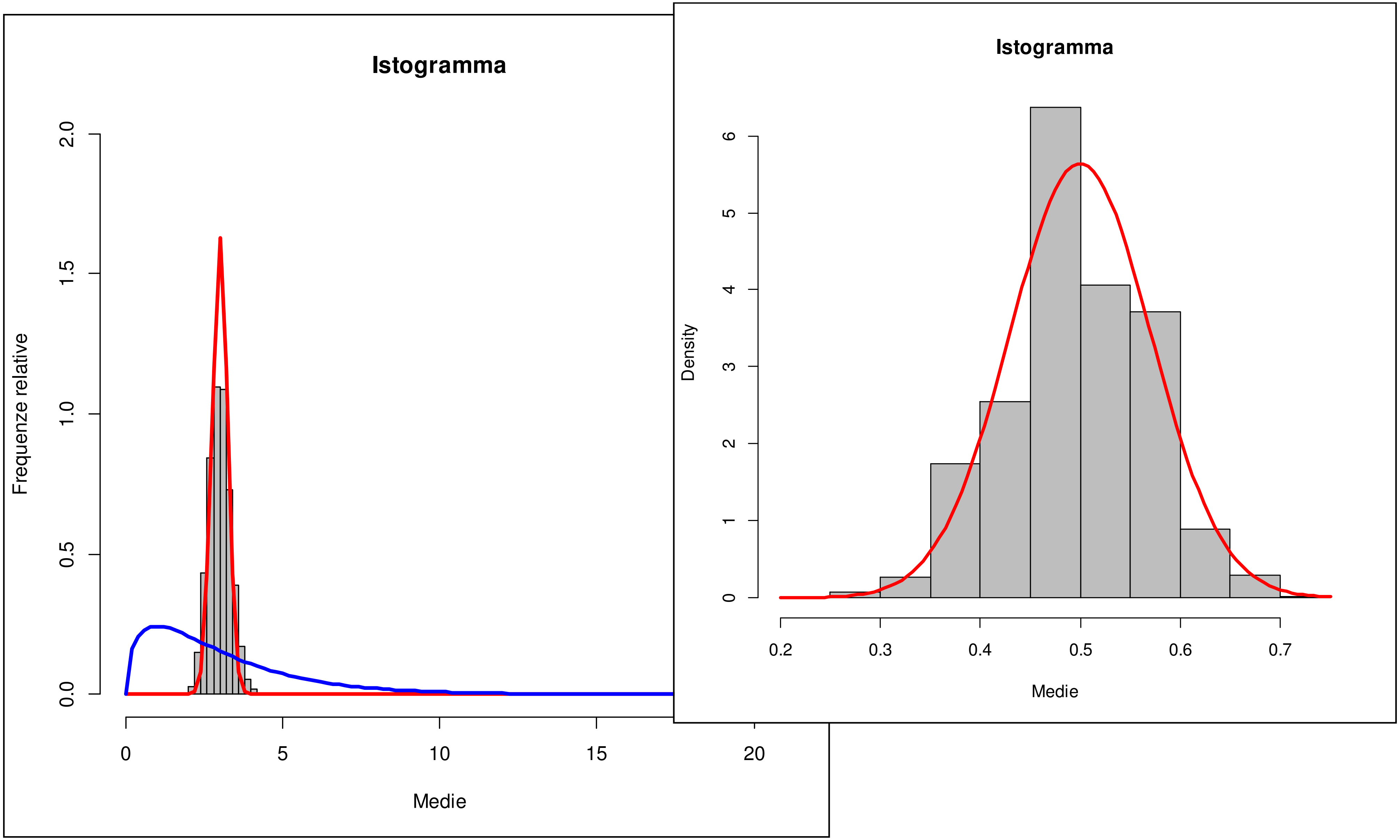
Se $X_i \sim N(\mu, \sigma^2)$ ⟹ $\bar{X}_n \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}^2}{\boldsymbol{n}}\right)$

Se $X_i \sim b(p)$ ⟹ $\bar{X}_n \sim \frac{1}{n} Bin(n, p)$

Teo. Centrale del Limite

Se $\boldsymbol{X_i \sim ???}$ **&** $\boldsymbol{n > 30}$ ⟹ $\bar{X}_n \sim \boldsymbol{N}\left(\boldsymbol{E(X_1)}, \frac{\boldsymbol{Var(X_1)}}{\boldsymbol{n}}\right)$

Facciamo un salto in R



Conoscendo la **distribuzione dello stimatore** conosciamo la **precisione della stima**

Riassunto

$X_1, X_2, \ldots, X_n$ i.i.d, $E(X_i) = \mu,$ $Var(X_i) = \sigma^2$

potenziali risultati che l'esperimento può dare

μ, σ^2 in generale **non noti**

$(x_1, x_2, \ldots, x_n) \Longrightarrow$ $\boxed{\bar{x}_n = \dfrac{x_1 + x_2 + \cdots + x_n}{n}}$

stima di μ

media campionaria

$\boxed{\bar{X}_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}}$

stimatore di μ

modello di tutte le possibili stime di μ, ***prima*** di estrarre il campione. Permette di dire **quanto è buona la stima** data da $\bar{x}_n$

Riassunto

$X_1, X_2, \dots, X_n$ i.i.d, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

media campionaria

$$\boxed{\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

stimatore di μ

modello di tutte le possibili stime di μ, ***prima*** di estrarre il campione. Permette di dire **quanto è buona la stima** data da $\bar{x}_n$

$$E(\bar{X}_n) = \mu$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Riassunto

$X_1, X_2, \ldots, X_n$ i.i.d, $E(X_i) = \mu,\quad Var(X_i) = \sigma^2$

media campionaria

$$\boxed{\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

stimatore di μ

$E(\bar{X}_n) = \mu$

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = n\mu$$

$$E(\bar{X}_n) = E\left(\frac{1}{n}[X_1 + X_2 + \cdots + X_n]\right) = \frac{1}{n} \times n\mu = \mu$$

Riassunto

$X_1, X_2, \ldots, X_n$ i.i.d, $E(X_i) = \mu,\qquad Var(X_i) = \sigma^2$

media campionaria

$$\boxed{\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

stimatore di μ

indipendenti

$E(\bar{X}_n) = \mu$

$$\boxed{Var(\bar{X}_n) = \frac{\sigma^2}{n}}$$

$$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$$

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}[X_1 + X_2 + \cdots + X_n]\right) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$$

Riassunto

$X_1, X_2, \dots, X_n$ i.i.d, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

media campionaria

$$\boxed{\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

stimatore di μ

$E(\bar{X}_n) = \mu$

$Var(\bar{X}_n) = \frac{\sigma^2}{n}$

Se σ^2 è noto, va a zero per campioni sempre più grandi ⇨ **più grande è il campione e maggiore è la <u>precisione</u> dello stimatore (e della stima).**

$$\boxed{\frac{n}{\sigma^2}}$$

Riassunto

$X_1, X_2, \dots, X_n$ i.i.d, $E(X_i) = \mu, \quad Var(X_i) = \sigma^2$

media campionaria

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

stimatore di μ $\quad \{ \ E(\bar{X}_n) = \mu \ \}$

varianza campionaria

$$\boxed{{S_n}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$

stimatore di σ^2

$$\{ \ E\left({S_n}^2\right) = \sigma^2 \ \}$$

Riassunto

$X_1, X_2, \dots, X_n$ i.i.d, $\sim N(\mu, \sigma^2)$
o $\boldsymbol{n}$ ***grande***

$$\boxed{E(X_i) = \mu, \qquad Var(X_i) = \sigma^2}$$

	MODELLO (stimatore)	DATI (stime)
media campionaria	$\bar{X}_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$	$\bar{x}_n = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
	stimatore di μ, $\bar{X}_n \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$	
varianza campionaria	$S_n{}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$	$s_n{}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
	stimatore di σ^2 $\{\ E(S_n{}^2) = \sigma^2\}$	$\dfrac{(n-1)}{\sigma^2} S_n{}^2 \sim \chi^2(n-1)$

Riassunto

$X_1, X_2, \dots, X_n$ i.i.d, $\sim \boldsymbol{b(p)}$

$\boldsymbol{P(X_i = 1) = p}$

$$\boxed{E(X_i) = p, \qquad Var(X_i) = p(1-p)}$$

	MODELLO (stimatore)	DATI (stime)
media campionaria	$\bar{X}_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$	$\hat{\boldsymbol{p}}_{\boldsymbol{n}} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
	stimatore di p, $E(\bar{X}_n) = p$	
varianza campionaria	$S_n{}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$	$s_n{}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{p}_n)^2$
	stimatore di $p(1-p)$ $\{ \ E\left(S_n{}^2\right) = p(1-p) \ \}$	$\boxed{\approx \hat{\boldsymbol{p}}_{\boldsymbol{n}}(\boldsymbol{1} - \hat{\boldsymbol{p}}_{\boldsymbol{n}})}$

Dal nostro *test*

2. Con la frase “fare inferenza sulla media” si intende:

Calcolare la media dei dati.	Stimare la media della popolazione da un campione casuale di dati.	NON SAPREI

Dal nostro *test*

2. Con la frase “fare inferenza sulla media” si intende:

Calcolare la media dei dati.	Stimare la media della popolazione da un campione casuale di dati.	NON SAPREI